

---

# Novel Bernstein-like Concentration Inequalities for the Missing Mass

---

**Bahman Yari Saeed Khanloo**  
 Monash University  
 bahman.khanloo@monash.edu

**Gholamreza Haffari**  
 Monash University  
 gholamreza.haffari@monash.edu

## Abstract

We are concerned with obtaining novel concentration inequalities for the *missing mass*, i.e. the total probability mass of the outcomes not observed in the sample. We not only derive - for the first time - distribution-free Bernstein-like deviation bounds with *sublinear* exponents in deviation size for missing mass, but also improve the results of McAllester and Ortiz (2003) and Berend and Kontorovich (2013, 2012) for small deviations which is the most interesting case in learning theory. It is known that the majority of standard inequalities cannot be directly used to analyze heterogeneous sums i.e. sums whose terms have large difference in magnitude. Our generic and intuitive approach shows that the heterogeneity issue introduced in McAllester and Ortiz (2003) is resolvable at least in the case of missing mass via regulating the terms using our novel thresholding technique.

## 1 INTRODUCTION

Missing mass is the total probability associated to the outcomes that have not been seen in the sample which is one of the important quantities in machine learning and statistics. It connects density estimates obtained from a given sample to the population for discrete distributions: the less the missing mass, the more useful the information that can be extracted from the dataset. Roughly speaking, the more the missing mass is the less we can discover about the true unknown underlying distribution which would imply the less we can statistically generalize to the whole population. In other words, missing mass measures how representative a given dataset is assuming that it has been sampled according to the true distribution.

Often, one is interested in understanding the behaviour of the missing mass as a random variable. One of the

important approaches in such studies involves bounding the fluctuations of the random variable around a certain quantity namely its mean. Concentration inequalities are powerful tools for performing analysis of this type. Let  $X$  be any non-negative real-valued random variable with finite mean. The goal is to establish for any  $\epsilon > 0$ , probability bounds of the form

$$\begin{aligned}\mathbb{P}(X - \mathbb{E}[X] \leq -\epsilon) &\leq \exp(-\eta_l(\epsilon)), \\ \mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) &\leq \exp(-\eta_u(\epsilon)),\end{aligned}\quad (1)$$

where  $\eta_l(\epsilon)$  and  $\eta_u(\epsilon)$  are some non-decreasing functions of  $\epsilon$  and where it is desirable to find the largest such functions for variable  $X$  and for the ‘target’ interval of  $\epsilon$ . These bounds are commonly called lower and upper deviations bounds respectively. In most practical scenarios, we are in a non-asymptotic setting where we have access to a sample  $X_1, \dots, X_n$  and we would like to derive concentration inequalities that explicitly describe dependence on sample size  $n$ . Namely, we would like to obtain bounds of the form

$$\begin{aligned}\mathbb{P}(X - \mathbb{E}[X] \leq -\epsilon) &\leq \exp(-\eta_l(\epsilon, n)), \\ \mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) &\leq \exp(-\eta_u(\epsilon, n)),\end{aligned}\quad (2)$$

where  $\eta_l(\epsilon, n)$  and  $\eta_u(\epsilon, n)$  are both non-decreasing functions of  $\epsilon$  and  $n$ . Many of such bounds are distribution-free i.e. they hold irrespective of the underlying distribution.

McAllester and Schapire (2000) established concentration inequalities for the missing mass for the first time. A follow-up work by McAllester and Ortiz (2003) pointed out inadequacy of standard inequalities, developed a thermodynamical viewpoint for addressing this issue and sharpened these bounds. Berend and Kontorovich (2013) further refined the bounds via arguments similar to Kearns-Saul inequality (Kearns and Saul (1998)) and logarithmic Sobolev inequality (Boucheron et al. (2013)). These previous works, however, not only involve overly specific approaches to concentration and handling heterogeneity issue but also do not yield sharp bounds for small deviations which is the most interesting case in learning theory.

In this paper, we shall derive distribution-free concentration inequalities for missing mass in a novel way. The

primary objective of our approach is to introduce a notion of *heterogeneity control* which allows us to *regulate* the magnitude of bins in histogram of the discrete distribution being analyzed. This mechanism in turn enables us to control the behaviour of central quantities such as the variance or martingale differences of the random variable in question. These are the main quantities that appear in standard concentration inequalities such as Bernstein, Bennett and McDiarmid just to name a few. Consequently, instead of discovering a new method for bounding fluctuations of each random variable of interest, we will be able to directly apply standard inequalities to obtain probabilistic bounds on many discrete random variables including missing mass.

The rest of the paper is structured as follows. Section 2 contains the background information and introduces the notations. Section 3 outlines motivations and the main contributions. In Section 4, we explain negative dependence, information monotonicity and develop a few fundamental tools whereas Section 5 presents the proofs of our upper and lower deviation bounds based on these tools. Finally, Section 6 concludes the paper and compares our bounds with existing results for small deviations.

## 2 PRELIMINARIES

In this section, we will provide definitions, notations and other background material.

Consider  $P : \mathcal{I} \rightarrow [0, 1]$  to be a fixed but unknown discrete distribution on some finite or countable non-empty set  $\mathcal{I}$  with  $|\mathcal{I}| = N$ . Let  $\{w_i : i \in \mathcal{I}\}$  be the probability (or frequency) of drawing the  $i$ -th outcome. Moreover, suppose that we observe an i.i.d. sample  $\{X_j\}_{j=1}^n$  from this distribution with  $n$  being the sample size. Now, missing mass is defined as the total probability mass corresponding to the outcomes that are not present in our sample. Namely, missing mass is a random variable that can be expressed as:

$$Y := \sum_{i \in \mathcal{I}} w_i Y_i, \quad (3)$$

where we define each  $\{Y_i : i \in \mathcal{I}\}$  to be a Bernoulli variable that takes on 0 if the  $i$ -th outcome exists in the sample and 1 otherwise. Namely, we have

$$Y_i = \mathbb{1}_{[(X_1 \neq i) \wedge (X_2 \neq i) \wedge \dots \wedge (X_n \neq i)]}. \quad (4)$$

We assume that for all  $i \in \mathcal{I}$ ,  $w_i > 0$  and  $\sum_{i \in \mathcal{I}} w_i = 1$ . Denote  $P(Y_i = 1) = q_i$  and  $P(Y_i = 0) = 1 - q_i$  and let us suppose that  $Y_i$ s are independent: as we will see later in this section, such an assumption will not impose a burden on our proof structure and flow. Hence, we will have that  $q_i = \mathbb{E}[Y_i] = (1 - w_i)^n \leq e^{-nw_i}$  where  $q_i \in (0, 1)$ . Namely, defining  $f : (1, n) \rightarrow (e^{-n}, \frac{1}{e}) \subset (0, 1)$  where  $f(\theta) = e^{-\theta}$  with  $\theta \in D_f$  and taking  $w_i > \frac{\theta}{n}$  amounts to  $q_i(w_i) \leq f(\theta)$ . This provides a basis for our ‘thresholding’ technique that we will employ in our proof.

Choosing the representation (3) for missing mass, one has

$$\mathbb{E}[Y]_{\mathcal{I}} = \sum_{i \in \mathcal{I}} w_i q_i = \sum_{i \in \mathcal{I}} w_i (1 - w_i)^n, \quad (5)$$

$$V[Y]_{\mathcal{I}} = \sum_{i \in \mathcal{I}} w_i^2 \text{VAR}[Y_i], \quad (6)$$

$$\underline{\sigma}_{\mathcal{I}}^2 := \sum_{i \in \mathcal{I}} w_i \text{VAR}[Y_i], \quad (7)$$

where we have introduced the weighted variance notation  $\underline{\sigma}^2$  and where each quantity is attached to a set over which it is defined. Note that  $\text{VAR}[Y_i]$  is the individual variance corresponding to  $Y_i$  which is defined as

$$\text{VAR}[Y_i] = q_i(1 - q_i) = (1 - w_i)^n (1 - (1 - w_i)^n). \quad (8)$$

One can define the above quantities not just over the set  $\mathcal{I}$  but on some (proper) subset of it that may depend on or be described by some variable(s) of interest. For instance, in our proofs the variable  $\theta$  may be responsible for choosing  $\mathcal{I}_{\theta} \subseteq \mathcal{I}$  over which the above quantities will be evaluated. For lower deviation and upper deviation, we find it convenient to refer to the associated set by  $\mathcal{L}$  and  $\mathcal{U}$  respectively. Likewise, we will use subscripts  $l$  and  $u$  to refer to objects that characterize lower deviation and upper deviation respectively. Also, we use the notation  $Y^{ij} = Y_i, \dots, Y_j$  to refer to sequence of variables whose index starts at  $i$ -th variable and ends at  $j$ -th variable. Finally, other notation or definitions may be introduced within the body of the proof when required.

We will encounter Lambert  $W$ -function - also known as product logarithm function - in this paper which describes the inverse relation of  $f(x) = xe^x$  and which cannot be expressed in terms of elementary functions. This function is double-valued when  $x \in \mathbb{R}$ . However, it becomes invertible in restricted domain. The lower branch of it is denoted by  $W_{-1}(\cdot)$ , which is the only branch that will prove beneficial in this paper. The reader is advised to refer to Corless et al. (1996) for a detailed treatment.

Throughout the paper, we shall use the convention that capital letters refer to random variables whereas lower case letters correspond to realizations thereof.

We will utilize Bernstein’s inequality in our derivation. Suitable representations of this result are outlined below without the proof.

**Theorem.** [Bernstein] *Let  $Z_1, \dots, Z_N$  be independent zero-mean random variables such that one has  $|Z_i| \leq \alpha$  almost surely for all  $i$ . Then, using Bernstein’s inequality (Bernstein (1924)) one obtains for all  $\epsilon > 0$ :*

$$\mathbb{P}\left(\sum_{i=1}^N Z_i > \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2(V + \frac{1}{3}\alpha\epsilon)}\right), \quad (9)$$

where  $V = \sum_{i=1}^N \mathbb{E}[Z_i^2]$ .

Now, consider the sample mean  $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$  and let  $\bar{\sigma}^2$  be the sample variance, namely  $\bar{\sigma}^2 := n^{-1} \sum_{i=1}^n \text{VAR}[Z_i] = n^{-1} \sum_{i=1}^n \mathbb{E}[Z_i^2]$ . So, using (9) with  $n \cdot \epsilon$  in the role of  $\epsilon$ , we get

$$\mathbb{P}(\bar{Z} > \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2(\bar{\sigma}^2 + \frac{1}{3}\alpha\epsilon)}\right). \quad (10)$$

If  $Z_1, \dots, Z_n$  are, moreover, not just independent but also identically distributed, then  $\bar{\sigma}^2$  is equal to  $\sigma^2$  i.e. the variance of each  $Z_i$ . The latter presentation makes explicit: (1) the exponential decay with  $n$ ; (2) the fact that for  $\bar{\sigma}^2 \leq \epsilon$  we get a tail probability with exponent of order  $n\epsilon$  rather than  $n\epsilon^2$  (Lugosi (2003); Boucheron et al. (2013)) which has the potential to yield stronger bounds for small  $\epsilon$ .

### 3 MOTIVATIONS AND MAIN RESULTS

In this section, we motivate this work by pointing out the heterogeneity challenge and how we approach it. Our bounds also improve the functional form of the exponent, which is of independent significance. In the final part of this section, we summarize our main results.

#### 3.1 The Challenge and the Remedy

McAllester and Ortiz (2003) point out that for highly heterogeneous sums of the form (3), the standard form of Bernstein's inequality (9) does not lead to concentration inequalities of form (10): at least for the upper deviation of the missing mass, (9) does not imply any non-trivial bounds of the form (2). The reason is basically the fact that the  $w_i$  can vary wildly: some can be of order  $O(1/n)$ , other may be constants independent of  $n$ . For similar reasons, other standard inequalities such as Bennett, Angluin-Valiant and Hoeffding cannot be used to get bounds on the missing mass of the form (2) either (McAllester and Ortiz (2003)).

Having pointed out the deficiency of these standard inequalities, McAllester and Ortiz (2003) succeed in giving bounds of the form (2) on the missing mass, for a function  $\eta(\epsilon, n) \propto n\epsilon^2$ , both with a direct argument and using the Kearns-Saul inequality (Kearns and Saul (1998)). Recently, the constants appearing in the bounds were refined by Berend and Kontorovich (2013). The bounds proven by McAllester and Ortiz (2003) and Berend and Kontorovich (2013) are qualitatively similar to Hoeffding bounds for i.i.d. random variables: they do *not* improve the functional form from  $n\epsilon^2$  to  $n\epsilon$  for small variances.

This leaves open the question whether it is also possible to derive bounds which are more reminiscent of the Bernstein bound for i.i.d. random variables (10) which does exploit variance. In this paper, we show that the answer is a qualified yes: we give bounds that depend on weighted variance  $\underline{\sigma}^2$  defined in (7) rather than sample

variance  $\bar{\sigma}^2$  as in (10) which is tight exactly in the important case when  $\underline{\sigma}^2$  is small, and in which the denominator in (10) is specified by a factor depending on  $\epsilon$ ; in the special case of the missing mass, this factor turns out to be logarithmic in  $\epsilon$  and a free parameter  $\gamma$  as it will become clear later.

We derive - using Bernstein's inequality - novel bounds on missing mass that take into account explicit variance information with more accurate scaling and demonstrate their superiority for small deviations.

#### 3.2 Main Results

Consider the following functions

$$\gamma_\epsilon = -2W_{-1}\left(-\frac{\epsilon}{2\sqrt{e}}\right), \quad (11)$$

$$c(\epsilon) = \frac{3(\gamma_\epsilon - 1)}{5\gamma_\epsilon^2}. \quad (12)$$

Let  $Y$  denote the missing mass,  $n$  the sample size and  $\epsilon$  the deviation size.

**Theorem 1.** *For any  $0 < \epsilon < 1$  and any  $n \geq \lceil \gamma_\epsilon \rceil - 1$ , we obtain the following upper deviation bound*

$$\mathbb{P}(Y - \mathbb{E}[Y] \geq \epsilon) \leq e^{-c(\epsilon) \cdot n\epsilon}. \quad (13)$$

**Theorem 2.** *For any  $0 < \epsilon < 1$  and any  $n \geq \lceil \gamma_\epsilon \rceil - 1$ , we obtain the following lower deviation bound*

$$\mathbb{P}(Y - \mathbb{E}[Y] \leq -\epsilon) \leq e^{-c(\epsilon) \cdot n\epsilon}. \quad (14)$$

**Corollary 1.** *For any  $0 < \epsilon < 1$  and any  $n \geq \lceil \gamma_\epsilon \rceil - 1$ , using union bound we obtain the following deviation bound*

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \epsilon) \leq 2e^{-c(\epsilon) \cdot n\epsilon}. \quad (15)$$

The proof of the above theorems is provided in Section 5. However, let us develop a few tools in Section 4 which will be used later in our proofs.

### 4 NEGATIVE DEPENDENCE AND INFORMATION MONOTONICITY

Probabilistic analysis of most random variables and specifically the derivation of the majority of probabilistic bounds rely on independence assumption between variables which offers considerable simplification and convenience. Many random variables including the missing mass, however, consist of random components that are not independent.

Fortunately, even in cases where independence does not hold, one can still use some standard tools and methods provided variables are dependent in specific ways. The following notions of dependence are among the common ways that prove useful in these settings: negative association and negative regression.

#### 4.1 Negative Dependence and Chernoff's Exponential Moment Method

Our proof involves variables with a specific type of dependence known as negative association. One can infer concentration of sums of negatively associated random variables from the concentration of sums of their independent copies in certain situations. In exponential moment method, this property allows us to treat such variables as independent in the context of probability inequalities as we shall elaborate later in this section.

In the sequel, we present negative association and regression and supply tools that will be essential in proofs.

**Negative Association:** Any real-valued random variables  $X_1$  and  $X_2$  are negatively associated if

$$\mathbb{E}[X_1 X_2] \leq \mathbb{E}[X_1] \cdot \mathbb{E}[X_2]. \quad (16)$$

More generally, a set of random variables  $X_1, \dots, X_m$  are negatively associated if for any disjoint subsets  $A$  and  $B$  of the index set  $\{1, \dots, m\}$ , we have

$$\mathbb{E}[X_i X_j] \leq \mathbb{E}[X_i] \cdot \mathbb{E}[X_j] \quad \text{for } i \in A, j \in B. \quad (17)$$

**Stochastic Domination:** Assume that  $X$  and  $Y$  are real-valued random variables. Then,  $X$  is said to stochastically dominate  $Y$  if for all  $a$  in the range of  $X$  and  $Y$  we have

$$P(X \geq a) \geq P(Y \geq a). \quad (18)$$

We use the notation  $X \succeq Y$  to reflect (18) in short.

**Stochastic Monotonicity:** A random variable  $Y$  is stochastically non-decreasing in random variable  $X$  if

$$x_1 \leq x_2 \implies P(Y|X = x_1) \leq P(Y|X = x_2). \quad (19)$$

Similarly,  $Y$  is stochastically non-increasing in  $X$  if

$$x_1 \leq x_2 \implies P(Y|X = x_1) \geq P(Y|X = x_2). \quad (20)$$

The notations  $(Y|X = x_1) \preceq (Y|X = x_2)$  and  $(Y|X = x_1) \succeq (Y|X = x_2)$  represent the above definitions using the notion of stochastic domination. Also, we will use shorthands  $Y \uparrow X$  and  $Y \downarrow X$  to refer to the relations described by (19) and (20) respectively.

**Negative Regression:** Random variables  $X$  and  $Y$  have negative regression dependence relation if  $X \downarrow Y$ .

Dubhashi and Ranjan (1998) as well as Joag-Dev and Proschan (1983) summarize numerous notable properties of negative association and negative regression. Specifically, the former provides a proposition that indicates that Hoeffding-Chernoff bounds apply to sums of negatively associated random variables. Further, McAllester and Ortiz (2003) generalize these observations to essentially any concentration result derived based on the

exponential moment method by drawing a connection between deviation probability of a discrete random variable and Chernoff's entropy of a related distribution.

We provide a self-standing account by presenting the proof for some of these existing results as well as developing several generic tools that are applicable beyond missing mass problem.

**Lemma 1. [Binary Stochastic Monotonicity]** Let  $Y$  be a binary random variable (Bernoulli) and let  $X$  take on values in a totally ordered set  $\mathcal{X}$ . Then, one has

$$Y \downarrow X \implies X \downarrow Y. \quad (21)$$

*Proof.* For any  $x$ , we have

$$\begin{aligned} P(Y = 1 | X \leq x) &\geq \inf_{a \leq x} P(Y = 1 | X = a) \\ &\geq \sup_{a > x} P(Y = 1 | X = a) \\ &\geq P(Y = 1 | X > x). \end{aligned} \quad (22)$$

The above argument implies that random variables  $Y$  and  $\mathbf{1}_{X > x}$  are negatively associated and since the expression  $P(X > x | Y = 1) \leq P(X > x | Y = 0)$  holds for all  $x \in \mathcal{X}$ , it follows that  $X \downarrow Y$ .  $\square$

**Lemma 2. [Independent Binary Negative Regression]**

Let  $X_1, \dots, X_m$  be negatively associated random variables and  $Y_1, \dots, Y_m$  be binary random variables (Bernoulli) such that either  $Y_i \downarrow X_i$  or  $Y_i \uparrow X_i$  holds for all  $i \in \{1, \dots, m\}$ . Then  $Y_1, \dots, Y_m$  are negatively associated.

*Proof.* For any disjoint subsets  $A$  and  $B$  of  $\{1, \dots, m\}$ , taking  $i \in A$  and  $j \in B$  we have

$$\mathbb{E}[Y_i Y_j] = \mathbb{E}[\mathbb{E}[Y_i Y_j | X_1, \dots, X_m]] \quad (23)$$

$$= \mathbb{E}[\mathbb{E}[Y_i | X_i] \cdot \mathbb{E}[Y_j | X_j]] \quad (24)$$

$$\leq \mathbb{E}[\mathbb{E}[Y_i | X_i]] \cdot \mathbb{E}[\mathbb{E}[Y_j | X_j]] \quad (25)$$

$$= \mathbb{E}[Y_i] \cdot \mathbb{E}[Y_j]. \quad (26)$$

Here, (24) holds since each  $Y_i$  only depends on  $X_i$ . Inequality (25) follows because  $X_i$  and  $X_j$  are negatively associated and we have  $\mathbb{E}[Y_i | X_i] = P(Y_i | X_i)$ .  $\square$

**Lemma 3. [Chernoff]** For any real-valued random variable  $X$  with finite mean  $\mathbb{E}[X]$  and for any  $x > 0$ , we have:

$$DP(X, x) \leq \exp(-S(X, x)), \quad (27)$$

$$S(X, x) = \sup_{\lambda} \{\lambda x - \ln(Z(X, \lambda))\}, \quad (28)$$

$$Z(X, \lambda) = \mathbb{E}[e^{\lambda X}]. \quad (29)$$

The lemma follows from the observation that for  $\lambda \geq 0$ , we have the following

$$P(X \geq x) = P(e^{\lambda X} \geq e^{\lambda x}) \leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda x}}. \quad (30)$$

This approach is known as *exponential moment method* (Chernoff (1952)) because of the inequality in (30).

**Lemma 4. [Negative Association]** In the exponential moment method, concentration of sums of negatively associated random variables can be deduced from the concentration of sums of their independent copies.

*Proof.* Let  $X_1, \dots, X_m$  be any set of negatively associated variables. Let  $X'_1, \dots, X'_m$  be independent shadow variables, i.e., independent variables such that each  $X'_i$  is distributed identically to  $X_i$ . Let  $X = \sum_i^m X_i$  and  $X' = \sum_i^m X'_i$ . For any set of negatively associated random variables, one has  $S(X, \epsilon) \geq S(X', \epsilon)$  since:

$$\begin{aligned} Z(X, \lambda) &= \mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[\prod_i^m e^{\lambda X_i}\right] \\ &\leq \prod_i^m \mathbb{E}[e^{\lambda X_i}] = \mathbb{E}[e^{\lambda X'}] = Z(X', \lambda). \end{aligned} \quad (31)$$

The lemma is due to McAllester and Ortiz (2003) which follows from definition of entropy function  $S$  given by (28).  $\square$

This lemma is very helpful in the context of large deviation bounds: it implies that one can treat negatively associated variables as if they were independent (McAllester and Ortiz (2003); Dubhashi and Ranjan (1998)).

**Lemma 5. [Balls and Bins]** Let  $S$  be any sample comprising  $n$  items drawn i.i.d. from a fixed distribution on integers  $\mathcal{N} = \{1, \dots, N\}$  (bins). Define  $C_i$  to be the number of times that integer  $i$  occurs in  $S$ . The random variables  $C_1, \dots, C_N$  are negatively associated.

*Proof.* Let  $f$  and  $g$  be non-decreasing and non-increasing functions respectively. We have

$$(f(x) - f(y))(g(x) - g(y)) \leq 0. \quad (32)$$

Further, assume that  $X$  is a real-valued random variable and  $Y$  is an independent shadow variable corresponding to  $X$ . Exploiting (32), we obtain

$$\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)] \cdot \mathbb{E}[g(X)], \quad (33)$$

which implies that  $f(X)$  and  $g(X)$  are negatively associated. Inequality (33) is an instance of Chebychev's fundamental *association inequality*.

Now, suppose without loss of generality that  $N = 2$ . Take  $X \in [0, n]$ , and consider the following functions

$$\begin{cases} f(X) = X, \\ g(X) = n - X, \end{cases} \quad (34)$$

where  $n = C_i + C_j$  is the total counts. Since  $f$  and  $g$  are non-decreasing and non-increasing functions of  $X$ , choosing  $X = f(C_i) = C_i$  we have for all  $i, j \in \mathcal{N}$  that

$$\mathbb{E}[C_i \cdot C_j] \leq \mathbb{E}[C_i] \cdot \mathbb{E}[C_j], \quad (35)$$

which concludes the proof for  $N = 2$ . Now, taking  $f(C_i) = C_i$  and  $g(C_i) = n - \sum_{j \neq i} C_j$  where  $n = \sum_{k=1}^N C_k$ , for  $N > 2$  the same argument implies that  $C_i$  and  $C_j$  are negatively associated for all  $i \in \mathcal{N}$  and  $j \in \mathcal{N} \setminus i$ . That is to say, any increase in  $C_i$  will cause a decrease in some or all of  $C_j$  variables with  $j \neq i$  and vice versa. It is easy to verify that the same is true for any disjoint subsets of the set  $\{C_1, \dots, C_N\}$ .  $\square$

**Lemma 6. [Monotonicity]** For any negatively associated random variables  $X_1, \dots, X_m$  and any non-decreasing functions  $f_1, \dots, f_m$ , we have that  $f_1(X_1), \dots, f_m(X_m)$  are negatively associated. The same holds if the functions  $f_1, \dots, f_m$  were non-increasing.

**Remark:** The proof is in the same spirit as that of association inequality (33) and motivated by composition rules for monotonic functions that one can repeatedly apply to (32).

**Lemma 7. [Union]** The union of independent sets of negatively associated random variables yields a set of negatively associated random variables.

Suppose that  $X$  and  $Y$  are independent vectors each of which comprising a negatively associated set. Then, the concatenated vector  $[X, Y]$  is negatively associated.

*Proof.* Let  $[X_1, X_2]$  and  $[Y_1, Y_2]$  be some arbitrary partitions of  $X$  and  $Y$  respectively and assume that  $f$  and  $g$  are non-decreasing functions. Then, one has

$$\begin{aligned} \mathbb{E}[f(X_1, Y_1) \cdot g(X_2, Y_2)] &= \\ \mathbb{E}[\mathbb{E}[f(X_1, Y_1) \cdot g(X_2, Y_2) \mid Y_1, Y_2]] &\leq \\ \mathbb{E}[\mathbb{E}[f(X_1, Y_1) \mid Y_1] \cdot \mathbb{E}[g(X_2, Y_2) \mid Y_2]] &\leq \\ \mathbb{E}[\mathbb{E}[f(X_1, Y_1) \mid Y_1]] \cdot \mathbb{E}[\mathbb{E}[g(X_2, Y_2) \mid Y_2]] &= \\ \mathbb{E}[f(X_1, Y_1)] \cdot \mathbb{E}[g(X_2, Y_2)]. \end{aligned} \quad (36)$$

The first inequality is due to independence of  $[X_1, X_2]$  from  $[Y_1, Y_2]$  which results in negative association being preserved under conditioning and the second inequality follows because  $[Y_1, Y_2]$  are negatively associated (Joag-Dev and Proschan (1983)). The same holds if  $f$  and  $g$  were non-increasing functions.  $\square$

**Lemma 8. [Splitting]** Splitting an arbitrary subset of bins of any fixed discrete distribution yields a set of negatively associated random bins.

*Proof.* Let  $w = (w_1, \dots, w_m)$  be a discrete distribution and  $\mathcal{W} = \{W_1, \dots, W_m\}$  be the associated set of random bins. Assume that  $w_i$  is split into  $k$  bins  $\mathcal{W}_i^S = \{W_{i1}, \dots, W_{ik}\}$  such that  $w_i = \sum_{j=1}^k W_{ij}$ . Then, by Lemma 5 members of split set  $\mathcal{W}_i^S$  are negatively associated. Clearly, the same holds for all  $1 \leq i \leq m$  as well as any other subset of set  $\mathcal{W}$ . Moreover, for all  $1 \leq i \leq m$  the sets  $\mathcal{W}_i^S$  and  $\mathcal{W} \setminus W_i$  are negatively associated by Lemma 5 and Lemma 7.  $\square$

**Lemma 9. [Absorption]** Absorbing any subset of bins of a discrete distribution yields negatively associated bins.

*Proof.* Let  $w = (w_1, \dots, w_N)$  be a discrete distribution and let  $\mathcal{W} = \{W_1, \dots, W_N\}$  be the associated set of random bins. Assume without loss of generality that  $\mathcal{W}^A = \{W_1^A, \dots, W_{N-1}^A\}$  is the absorption-induced set of random bins where  $w_N$  is absorbed to produce  $w^A = (w_1^A, \dots, w_{N-1}^A)$  and where  $w_i^A = w_i + \frac{w_N}{N-1}$  for  $i = 1, \dots, N-1$ . So,  $w_N$  is discarded and we have  $\sum_{i=1}^{N-1} w_i^A = 1 - w_N$ . The rest of the proof concerns applying Lemma 5 to the absorb set  $\mathcal{W}^A$ . The same holds if we absorb  $w_N$  to a subset of  $\mathcal{W} \setminus W_N$ .  $\square$

## 4.2 Negative Dependence and the Missing Mass

For missing mass, the variables  $W_i = \frac{C_i}{n}$  are negatively associated owing to Lemma 5 and linearity of expectation. Also, one has  $\forall i : Y_i \downarrow W_i$ . So, by Lemma 1 we infer that  $\forall i : W_i \downarrow Y_i$ . Now,  $Y_1, \dots, Y_N$  are negatively associated because they are a set of independent binary variables with negative regression dependence (Lemma 2). Thus, concentration variables

$Z_i = w_i Y_i - \mathbb{E}[w_i Y_i] := \zeta(Y_i)$  are negatively associated by Lemma 6 since we have

$$\zeta(Y_i) = \begin{cases} -w_i q_i & \text{if } Y_i = 0, \\ w_i(1 - q_i) & \text{if } Y_i = 1. \end{cases} \quad (37)$$

For all  $i$ ,  $\zeta$  is a non-decreasing function of  $Y_i$ . Likewise, concentration variables  $-Z_i$  are

negatively associated.

## 4.3 Information Monotonicity and Partitioning

**Lemma 10. [Information Monotonicity]** Let  $p = (p_1, \dots, p_N)$  be a discrete distribution on  $X = (x_1, \dots, x_N)$  such that for  $1 \leq i \leq N$  we have  $P(X = x_i) = p_i$ . Suppose we partition  $X$  into  $m \leq N$  non-empty disjoint groups  $G_1, \dots, G_m$ , namely

$$X = \cup G_i, \quad \forall i \neq j : G_i \cap G_j = \emptyset. \quad (38)$$

This is called *coarse binning* since it generates a new distribution with groups  $G_i$  whose dimensionality is less than that of the original distribution. Note that once the distribution is transformed, considering any outcome  $x_i$  from the original distribution we will only have access to its group membership information; for instance, we can observe that it belongs to  $G_j$  but we will not be able to recover  $p_i$ .

Let us denote the induced distribution over the partition  $G = (G_1, \dots, G_m)$  by  $p^G = (p_1^G, \dots, p_m^G)$ . Clearly, we have

$$p_i^G = P(G_i) = \sum_{j \in G_i} P(x_j). \quad (39)$$

Now, consider the  $f$ -divergence  $D_f(p^G \| q^G)$  between induced probability distributions  $p^G$  and  $q^G$ . Information monotonicity states that information is lost as we partition elements of  $p$  and  $q$  into groups to produce  $p^G$  and  $q^G$  respectively. Namely, for any  $f$ -divergence one has

$$D_f(p^G \| q^G) \leq D_f(p \| q), \quad (40)$$

which is due to Csiszár (Csiszár (1977, 2008); Amari (2009)). This inequality is tight if and only if for any outcome  $x_i$  and partition  $G_j$ , we have  $p(x_i | G_j) = q(x_i | G_j)$ .

**Lemma 11. [Partitioning]** In the exponential moment method, one can establish a deviation bound for any discrete random variable  $X$  by invoking Chernoff's method on the associated discrete partition random variable  $X^G$ .

Formally, assume  $X$  and  $X_\lambda$  are discrete random variables defined on the set  $\mathcal{X}$  endowed with probability distributions  $p$  and  $p_\lambda$  respectively. Further, suppose that  $X^G$  and  $X_\lambda^G$  are discrete variables on a partition set  $\mathcal{X}^G$  endowed with  $p^G$  and  $p_\lambda^G$  that are obtained from  $p$  and  $p_\lambda$  by partitioning using some partition  $G$ . Then, we have

$$\forall x > 0 : DP(X, x) \leq \exp(-S(X^G, x)). \quad (41)$$

*Proof.* Let  $\lambda(x)$  be the optimal  $\lambda$  in (28). Then, we have

$$\begin{aligned} S(X, x) &= x\lambda(x) - \ln(Z(X, \lambda(x))) \\ &= D_{KL}(p_{\lambda(x)} \| p) \\ &\geq D_{KL}(p_{\lambda(x)}^G \| p^G) \\ &= S(X^G, x), \end{aligned} \quad (42)$$

where we have introduced the  $\lambda$ -induced distribution

$$P_\lambda(X = x) = \frac{e^{\lambda x}}{Z(X, \lambda)} P(X = x). \quad (43)$$

The inequality step in (42) follows from (40) and the observation that  $D_{KL}$  is an instance of  $f$ -divergence where  $f(v) = v \ln(v)$  with  $v \geq 0$ .  $\square$

## 5 PROOF OF THE MAIN RESULTS

The central idea of the proof is to regulate the terms in the sum given by (3) via controlling the magnitude of bins of the distribution using operations that preserve negative association. This mechanism will help defeat the heterogeneity issue leading to the failure of standard probability inequalities described by McAllester and Ortiz (2003).

### 5.1 Proof of Theorem 1: Upper Deviation Bound

We consider the thresholds  $\tau = \frac{\theta}{n}$  and  $\tau' = \frac{2\theta}{n}$  and reduce the problem to one in which all bins that are larger than  $\tau$  are eliminated, where  $\theta \in \mathbb{R}$  will depend on the target deviation size  $\epsilon$ .

The reduction is performed by *splitting* the bins that are larger than  $\tau$  and then *absorbing* the bins that are smaller than  $\tau$ . This is followed by choosing a threshold that yields the sharpest bound for the choice of  $\epsilon$ . It turns out that the optimal threshold will too be a function of  $\epsilon$ .

Let  $\mathcal{I}_\tau \subseteq \mathcal{I}$  denote the subset of bins that are at most as large as  $\tau$ ,  $\mathcal{I}_\theta$  the subset of bins whose magnitude is between  $\tau$  and  $\tau'$ ,  $\mathcal{I}_{\tau'}$  the subset of bins larger than  $\tau'$  and  $\mathcal{I}'_\theta$  and  $\mathcal{I}'_{\tau'}$  the set of bins that we obtain after splitting members of  $\mathcal{I}_\theta$  and  $\mathcal{I}_{\tau'}$  respectively.

Now, for each  $i \in \mathcal{I} \setminus \mathcal{I}_\tau = \{\mathcal{I}_\theta \cup \mathcal{I}_{\tau'}\}$  and for some  $k \in \mathbb{N}$  that depends on  $i$  (but we suppress that notation below), we will have that  $k \cdot \tau \leq w_i < (k+1) \cdot \tau$ . For all such  $i$ , we define extra independent Bernoulli random variables  $Y_{ij}$  with  $j \in \mathcal{J}_i := \{1, \dots, k\}$  and their associated bins  $w_{ij}$ . For  $j \in \{1, \dots, k-1\}$ ,  $w_{ij} = \tau$  and  $w_{ik} = w_i - (k-1) \cdot \tau$ . In this way, all bins that are larger than  $\tau$  are split up into  $k$  bins, each of which is in-between  $\tau$  and  $\tau'$ ; more precisely, the first  $k-1$  are exactly  $\tau$  and the last one may be larger. Therefore, we consider the split random variable  $Y' = \sum_{i \in \mathcal{I}_\tau} w_i Y_i + \sum_{i \in \{\mathcal{I}'_\theta \cup \mathcal{I}'_{\tau'}\}} \sum_{j \in \mathcal{J}_i} w_{ij} Y_{ij}$  and the set  $\mathcal{U}' = \{i \mid w_i < \tau'\} = \{\mathcal{I}_\tau \cup \mathcal{I}'_\theta \cup \mathcal{I}'_{\tau'}\}$ . Furthermore, we introduce the random variable  $Y'' = \sum_{i \in \mathcal{U}''} w_i Y_i$  on the absorption-induced set  $\mathcal{U}'' = \{i \mid \tau \leq w_i < \tau'\}$ . The set  $\mathcal{U}''$  is generated from  $\mathcal{U}'$  as follows: we take the largest element  $j \in \mathcal{U}'$  with  $w_j < \tau$ , update  $w_l$  using  $w_l \leftarrow w_l + \frac{w_j}{|\mathcal{U}'|-1}$  for  $\{l \in \mathcal{U}' : l \neq j, w_l < \tau\}$  and discard  $w_j$ . Repeating this procedure gives a set of bins whose sizes are in-between  $\tau$  and  $\tau'$  plus a single bin of size smaller than  $\tau$ ; absorbing the latter into one of the members of the former with size  $\tau$  yields  $\mathcal{U}''$ .

Now, by choosing  $\theta$  such that  $f(\theta) = e^{-\theta} = \frac{\epsilon}{\gamma}$  and  $\theta = f^{-1}(\frac{\epsilon}{\gamma}) = \ln(\frac{\gamma}{\epsilon})$  for any  $0 < \epsilon < 1$  and  $e\epsilon < \gamma < e^n \epsilon$  as generic domain for  $\gamma$ , we derive the upper deviation bound for missing mass as follows

$$\mathbb{P}(Y - \mathbb{E}[Y] \geq \epsilon) \leq \quad (44)$$

$$\mathbb{P}(Y' - \mathbb{E}[Y] \geq \epsilon) = \quad (45)$$

$$\mathbb{P}(Y' - \mathbb{E}[Y'] + (\mathbb{E}[Y'] - \mathbb{E}[Y]) \geq \epsilon) \leq \quad (46)$$

$$\mathbb{P}(Y' - \mathbb{E}[Y'] + f(\theta) \geq \epsilon) = \quad (47)$$

$$\mathbb{P}\left(Y' - \mathbb{E}[Y'] \geq \left(\frac{\gamma-1}{\gamma}\right)\epsilon\right) = \quad (48)$$

$$\exp\left(-\frac{\left(\frac{\gamma-1}{\gamma}\right)^2 \epsilon^2}{2(V_{\mathcal{U}''} + \frac{\alpha_n}{3} \cdot \left(\frac{\gamma-1}{\gamma}\right) \cdot \epsilon)}\right) \leq \quad (49)$$

$$\exp\left(-\frac{\left(\frac{\gamma-1}{\gamma}\right)^2 \epsilon^2}{2\left(\frac{\theta}{n} \cdot \epsilon + \frac{2\theta}{3n} \cdot \left(\frac{\gamma-1}{\gamma}\right) \cdot \epsilon\right)}\right) \leq \quad (50)$$

$$\inf_{\gamma} \left\{ \exp\left(-\frac{3n\epsilon(\gamma-1)^2}{10\gamma^2 \ln(\frac{\gamma}{\epsilon})}\right) \right\} = \quad (51)$$

$$e^{-c(\epsilon) \cdot n\epsilon}. \quad (52)$$

Clearly, we will have that  $\tau^* = \frac{\theta^*}{n}$  where  $\theta^* = \ln(\frac{\gamma}{\epsilon})$ .

Inequality (45) follows because the splitting procedure cannot decrease deviation probability of missing mass.

Formally, assume without loss of generality that  $\mathcal{I} \setminus \mathcal{I}_\tau$  has only one element corresponding to  $Y_1$ ,  $\mathcal{J}_1 = \{1, 2\}$  and  $k_1 = 1$  i.e.  $w_1$  is split into two parts. Then, deviation probability of  $Y$  can be thought of as the total probability mass associated to independent Bernoulli variables  $Y_1, \dots, Y_N$  whose weighted sum is bounded below by some tail  $t > 0$ . Hence, we have

$$\begin{aligned} \mathbb{P}(Y \geq t) &= \sum_{Y^{1N}; \hat{Y} \geq t} P(Y_1, \dots, Y_N) \\ &= \sum_{Y^{1N}; \hat{Y} \geq t} R(Y_1) \cdot \prod_{i=2}^N R(Y_i) \\ &\quad + \sum_{Y^{1N}; \hat{Y} < t; Y \geq t} R(Y_1) \cdot \prod_{i=2}^N R(Y_i) \\ &= \sum_{Y^{1N}; \hat{Y} \geq t} R(Y_1) \cdot \prod_{i=2}^N R(Y_i) \\ &\quad + \sum_{Y^{1N}; \hat{Y} < t; Y \geq t, Y_1=1} R(Y_1) \cdot \prod_{i=2}^N R(Y_i) \\ &= \sum_{Y^{2N}; \hat{Y} \geq t} \prod_{i=2}^N R(Y_i) \\ &\quad + \sum_{Y^{2N}; \hat{Y} < t; Y \geq t} q_1 \cdot \prod_{i=2}^N R(Y_i), \end{aligned} \quad (53)$$

where  $\hat{Y} = \sum_{i \geq 2} w_i Y_i$  and  $R(Y_i) = q_i$  if  $Y_i = 1$  and  $R(Y_i) = 1 - q_i$  otherwise. Likewise, one can express the upper deviation probability of  $Y'$  as follows

$$\begin{aligned} \mathbb{P}(Y' \geq t) &= \sum_{Y^{1N}; \hat{Y} \geq t} R(Y_1) \cdot \prod_{i=2}^N R(Y_i) \\ &\quad + \sum_{Y^{11}, Y^{12}, Y^{2N}; \hat{Y} < t; Y' \geq t} \left( R(Y_{11}) \cdot R(Y_{12}) \right) \prod_{i=2}^N R(Y_i) \\ &= \sum_{Y^{2N}; \hat{Y} \geq t} \prod_{i=2}^N R(Y_i) \\ &\quad + \sum_{Y^{11}, Y^{12}, Y^{2N}; \hat{Y} < t; Y' \geq t} \left( R(Y_{11}) \cdot R(Y_{12}) \right) \prod_{i=2}^N R(Y_i) \\ &\geq \sum_{Y^{2N}; \hat{Y} \geq t} \prod_{i=2}^N R(Y_i) \\ &\quad + \sum_{Y^{2N}; \hat{Y} < t; Y' \geq t} (q_{11} \cdot q_{12}) \prod_{i=2}^N R(Y_i), \end{aligned} \quad (54)$$

where  $R(Y_{ij}) = q_{ij}$  if  $Y_{ij} = 1$  and  $R(Y_{ij}) = 1 - q_{ij}$  otherwise. Thus, combining (53) and (54) we have

$$\begin{aligned} \mathbb{P}(Y' \geq t) - \mathbb{P}(Y \geq t) &\geq \\ \sum_{Y^{2N}; \hat{Y} < t; Y' \geq t; Y \geq t} (q_{11} \cdot q_{12} - q_1) \prod_{i=2}^N R(Y_i) &= \\ \sum_{Y^{2N}; \hat{Y} < t; Y' \geq t} (q_{11} \cdot q_{12} - q_1) \prod_{i=2}^N R(Y_i). \end{aligned} \quad (55)$$

To complete the proof for (45), we require the expression for the difference between deviation probabilities in (55) to be non-negative for all  $t > 0$  which holds if  $q_1 \leq q_{11} \cdot q_{12}$ . For the missing mass, this condition holds. Without loss of generality, assume that  $w_i$  is split into two terms; namely, we have  $w_i = w_{ij} + w_{ij'}$ . Then, we can check the above condition as follows

$$\begin{aligned} q_i &= (1 - w_i)^n \leq (1 - w_{ij})^n \cdot (1 - w_{ij'})^n \\ &= \left(1 - \underbrace{(w_{ij} + w_{ij'})}_{w_i} + \underbrace{w_{ij} \cdot w_{ij'}}_{\geq 0}\right)^n. \end{aligned} \quad (56)$$

One can verify using induction that (56) holds also for cases where the split operation produces more than two terms. Now, choosing tail size  $t = \epsilon + \mathbb{E}Y$  implies (45).

Inequality (47) follows because the gap between the expectations will be negligible. Denoting  $\mathbb{E}[Y'_i] = q'_i$ , we have

$$q'_i = \begin{cases} q_i & \text{if } i \in \mathcal{I}_\tau, \\ q_{ij} & \text{if } i \in \{\mathcal{I}'_\tau, \cup \mathcal{I}'_\theta\}, \\ 0 & \text{otherwise.} \end{cases} \quad (57)$$

Namely, we can write

$$\begin{aligned} g_u(\theta) &= \mathbb{E}[Y'] - \mathbb{E}[Y] = \sum_{i \in \mathcal{I}} w_i (q'_i - q_i) \\ &= \sum_{i \in \mathcal{I}_\tau} w_i q_i + \sum_{i \in \{\mathcal{I}'_\tau, \cup \mathcal{I}'_\theta\}} \sum_{j \in \mathcal{J}_i} w_{ij} q_{ij} - \sum_{i \in \mathcal{I}} w_i q_i \\ &= \sum_{i \in \{\mathcal{I}'_\tau, \cup \mathcal{I}'_\theta\}} \sum_{j \in \mathcal{J}_i} w_{ij} q_{ij} - \sum_{i \in \{\mathcal{I}_\tau, \cup \mathcal{I}_\theta\}} w_i q_i \\ &\leq \sum_{i \in \{\mathcal{I}'_\tau, \cup \mathcal{I}'_\theta\}} \sum_{j \in \mathcal{J}_i} w_{ij} q_{ij} \\ &\leq \sum_{i \in \{\mathcal{I}'_\tau, \cup \mathcal{I}'_\theta\}} \sum_{j \in \mathcal{J}_i} w_{ij} f(\theta) \leq f(\theta). \end{aligned} \quad (58)$$

The expression in (49) is Bernstein's inequality applied to the random variable  $Z_u = \sum_{i \in \mathcal{U}''} Z_i$  relying upon Lemma 11. Here, the concentration variables are  $Z_i = w_i Y_i - \mathbb{E}[w_i Y_i]$  with  $i \in \mathcal{U}''$  and we set  $\alpha_u = \tau'$ .

Let  $V_{\mathcal{U}''}$  be variance proxy term  $V$  in Bernstein's inequality as defined in (9) attached to  $\mathcal{U}''$ . The functions  $f, g : (0, 1) \times \mathbb{N} \rightarrow (0, 1)$  with  $f(x, n) = x(1-x)^n(1-(1-x)^n)$

and  $g(x, n) = x^2(1-x)^n(1-(1-x)^n)$  are non-increasing with respect to  $x$  on  $(\frac{1}{n+1}, 1)$  and  $(\frac{2}{n+2}, 1)$  respectively. We obtain for  $1 < \theta < n$ , an upperbound on  $V_{\mathcal{U}''}$  as follows:

$$\begin{aligned} V_{\mathcal{U}''} &= \sum_{i: w_i \in \mathcal{U}''} w_i^2 (1 - w_i)^n \left(1 - (1 - w_i)^n\right) \\ &\leq \tau \cdot \sum_{i: w_i \in \mathcal{U}''} w_i (1 - w_i)^n \left(1 - (1 - w_i)^n\right) \\ &= \tau \cdot \underline{\sigma}_{\mathcal{U}''}^2 \\ &\leq \tau \cdot \sum_{i: \tau \leq w_i < \tau'; \sum_i w_i = 1} w_i (1 - w_i)^n \\ &\leq \underbrace{|\mathcal{I}_{(\theta, n)}|}_{\leq \frac{n}{\theta}} \cdot \left(\frac{\theta}{n}\right)^2 \cdot \left(1 - \frac{\theta}{n}\right)^n \\ &\leq \frac{\theta}{n} \cdot e^{-\theta} < \frac{\theta}{n} \cdot \epsilon. \end{aligned} \quad (59)$$

In order to see why (52) holds, consider  $c(\gamma, \epsilon) = \frac{\epsilon(\gamma-1)^2}{\gamma^2 \ln(\frac{\gamma}{\epsilon})}$  and let us examine the derivatives as follows

$$\frac{\partial c(\gamma, \epsilon)}{\partial \gamma} = -\frac{\epsilon^2(\gamma-1)(\gamma-1-2\ln(\frac{\gamma}{\epsilon}))}{\gamma^3 \ln^2(\frac{\gamma}{\epsilon})}, \quad (60)$$

$$\begin{aligned} \frac{\partial^2 c(\gamma, \epsilon)}{\partial \gamma^2} &= \frac{\epsilon^2}{\gamma^4 \ln^3(\frac{\gamma}{\epsilon})} \left[ (6-4\gamma) \ln^2\left(\frac{\gamma}{\epsilon}\right) + \right. \\ &\quad \left. (\gamma^2 - 6\gamma + 5) \ln\left(\frac{\gamma}{\epsilon}\right) + 2(\gamma-1)^2 \right]. \end{aligned} \quad (61)$$

Solving for the first derivative using (60), we obtain

$$\gamma_\epsilon = -2W_{-1}\left(-\frac{\epsilon}{2\sqrt{e}}\right). \quad (62)$$

Inspecting the second derivative given by (61), we can see that the function  $c(\gamma, \epsilon)$  is concave with respect to  $\gamma$  for any  $\gamma > 2$ . Recall, moreover, that there are interrelated restrictions on  $\gamma$ ,  $\epsilon$  and  $n$  in derivation of (51) and (52) which are collectively expressed as

$$\max\{e \cdot \epsilon, 1, 2, \gamma(1)\} < \gamma < e^n, \quad n \geq \lceil \gamma_\epsilon \rceil - 1. \quad (63)$$

## 5.2 Proof of Theorem 2: Lower Deviation Bound

The proof for lower deviation bound proceeds in the same spirit as section 5.1. The idea is again to reduce the problem to one in which all bins that are larger than the threshold  $\tau$  are eliminated.

We split large bins and then absorb small bins to enable us shrink the variance while controlling the magnitude of terms (and consequently the key quantities  $\alpha$  and  $V$ ) before applying Bernstein's inequality.

By choosing  $\theta$  such that  $f(\theta) = e^{-\theta}$  so that  $\theta = \ln(\frac{\gamma}{\epsilon})$ , for any  $0 < \epsilon < 1$  with  $e\epsilon < \gamma < e^n \epsilon$  being generic domain



for  $\gamma$  we obtain a lower deviation bound as follows

$$\mathbb{P}(Y - \mathbb{E}[Y] \leq -\epsilon) \leq \quad (64)$$

$$\mathbb{P}(Y' - \mathbb{E}[Y] \leq -\epsilon) = \quad (65)$$

$$\mathbb{P}(Y' - \mathbb{E}[Y'] + (\mathbb{E}[Y'] - \mathbb{E}[Y]) \leq -\epsilon) \leq \quad (66)$$

$$\mathbb{P}(Y' - \mathbb{E}[Y'] - f(\theta) \leq -\epsilon) = \quad (67)$$

$$\mathbb{P}\left(Y' - \mathbb{E}[Y'] \leq -\left(\frac{\gamma-1}{\gamma}\right)\epsilon\right) \leq \quad (68)$$

$$\leq \exp\left(-\frac{\left(\frac{\gamma-1}{\gamma}\right)^2 \epsilon^2}{2(V_{\mathcal{L}''} + \frac{\alpha_l}{3} \cdot \left(\frac{\gamma-1}{\gamma}\right) \cdot \epsilon)}\right) \leq \quad (69)$$

$$\leq \exp\left(-\frac{\left(\frac{\gamma-1}{\gamma}\right)^2 \epsilon^2}{2\left(\frac{\theta}{n} \cdot \epsilon + \frac{2\theta}{3n} \cdot \left(\frac{\gamma-1}{\gamma}\right) \cdot \epsilon\right)}\right) \leq \quad (70)$$

$$\inf_{\gamma} \left\{ \exp\left(-\frac{3n\epsilon(\gamma-1)^2}{10\gamma^2 \ln(\frac{2}{\epsilon})}\right) \right\} = \quad (71)$$

$$e^{-c(\epsilon) \cdot n\epsilon}, \quad (72)$$

where  $c(\epsilon)$  and  $\tau^*$  are as before and domain restrictions are determined similar to (63).

The variables  $Y'$  and  $Y''$ , and the sets  $\mathcal{L}'$  and  $\mathcal{L}''$  are defined in the same fashion as Section 5.1.

The first inequality is proved in the same way as (45). Now, we set  $\mathbb{E}[Y'_i] = q'_i$  such that

$$q'_i = \begin{cases} q_i & \text{if } w_i < \tau', \\ 0 & \text{otherwise.} \end{cases} \quad (73)$$

Inequality (67) follows because the compensation gap will remain small since we have

$$\begin{aligned} g_l(\theta) &= \mathbb{E}[Y'] - \mathbb{E}[Y] = \sum_{i \in \mathcal{I}} w_i(q'_i - q_i) \\ &= \sum_{i: w_i < \tau'} w_i q_i - \sum_{i \in \mathcal{I}} w_i q_i = - \sum_{i: w_i \geq \tau'} w_i q_i \\ &\geq - \sum_{i: w_i \geq \tau'} w_i f(\theta) \geq -f(\theta). \end{aligned} \quad (74)$$

The expression given by (69) is Bernstein's inequality applied to random variable  $Z_l = \sum_{i \in \mathcal{L}''} Z_i$  where we have defined  $Z_i = w_i(\mu - w_i Y_i) - \mathbb{E}[w_i(\mu - w_i Y_i)]$  with  $\mu$  being the upper bound on the value of the  $w_i Y_i$  terms.

Further, we choose  $\alpha_l = \tau'$ . Observe that  $Z_l = -Z_u$  and  $\mu = \alpha_l$ . Finally, an upperbound on  $V_{\mathcal{L}''}$  can be determined with arguments identical to that of  $V_{\mathcal{L}''}$ .

The rest of the proof proceeds in an analogous manner to the proof of upper deviation bound.

## 6 CONCLUSIONS

We proposed a new technique for establishing concentration inequalities and applied it to the missing mass using

Bernstein's inequality. Along the way, we introduced a collection of concepts and tools in the intersection of probability theory and information theory that have the potential to be advantageous in more general settings.

Recall that Bernstein's inequality hinges on establishing an upperbound on  $Z(X, \lambda)$  given by (29) in a particular way. Clearly, this choice is not unique and one can choose any other upperbound (e.g. c.f. Lugosi (2003)) and apply the same technique to derive potentially tight bounds achievable within the framework of exponential moment method.

Our bounds sharpen the leading results for missing mass in the case of small deviations. These inequalities hold subject to the mild condition that the sample size is large enough, namely  $n \geq \lceil \gamma_\epsilon \rceil - 1$ .

We select the best known bounds in Berend and Kontorovich (2013) for the comparison. Our lower deviation and upper deviation bounds improve state-of-the-art for any  $0 < \epsilon < 0.021$  and any  $0 < \epsilon < 0.045$  respectively.

Plugging in the definitions, we can see that the compensation gap can be expressed as a function of  $\epsilon$  and show that the following holds

$$|g(\epsilon)| \leq \sqrt{e} \cdot \exp\left(W_{-1}\left(\frac{-\epsilon}{2\sqrt{e}}\right)\right), \quad (75)$$

where we have dropped the subscript of gap  $g$ . Note that the gap is negligible for small  $\epsilon$  compared to large values of  $\epsilon$  for both (52) and (72). This observation supports the fact that we obtained sharper bounds for small deviations.

Mathematical analysis of missing mass via concentration inequalities has various important applications including density estimation, generalization bounds and handling missing data just to name a few. Needless to say that any refinement in bounds or tools developed for the former may directly contribute to advancement in those applications.

## Acknowledgements

The authors are grateful to National ICT Australia (NICTA) for generous funding, as part of collaborative machine learning research projects. We would like to thank Peter Grünwald, Aryeh Kontorovich, Thijs van Ommen and Mark Schmidt for feedback and helpful discussions, and anonymous reviewers for their constructive comments.

## References

- Shun-ichi Amari. Divergence function, information monotonicity and information geometry. *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE)*, 2009.
- Daniel Berend and Aryeh Kontorovich. The missing mass problem. *Statistics & Probability Letters*, 2012.

- Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18:no. 3, 1–7, 2013.
- S. N. Bernstein. On a modification of Chebyshev’s inequality and of the error formula of Laplace. *Annals Science Institute SAV. Ukraine*, 1924.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- H. Chernoff. A measure of the asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.
- R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W function. In *Advances in Computational Mathematics*, 1996.
- Imre Csiszár. Information measures: a critical survey. *7th Prague Conference on Information Theory*, pages 73–86, 1977.
- Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10:261–273, 2008.
- Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13:99–124, 1998.
- Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *Annals of Statistics*, 11:286–295, 1983.
- Michael Kearns and Lawrence Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 1998.
- Gábor Lugosi. Concentration of measure inequalities, 2003. URL <http://www.econ.upf.es/~lugosi/anu.ps>.
- David McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research (JMLR)*, 4, 2003.
- David McAllester and Robert E. Schapire. On the convergence rate of Good-Turing estimators. *Conference on Learning Theory (COLT)*, 2000.
- Robin Pemantle. Towards a theory of negative dependence. *Journal of Mathematical Physics*, 41:1371–1390, 1999.